胡杨8月工作日志（2014.8.29）

1. 不良产品信息全文检索功能增强
2. 为不良产品全文检索增加检索关键词高亮显示及自动摘要功能，前台显示功能由戴丹实现。
3. 采集器功能完善与增强
4. 采集器增加对新浪微博中热门微博及其评论的采集功能，实现对于新浪微博的模拟登陆和内容解析。
5. 分类器，采集器性能调优及错误修改
6. 采集器，grab部分bug修改，AnalyzerTime时间处理类bug修改。
7. 分类器，添加LibSVM支持，LibSVM是优秀的开源SVM算法实现，用于替换weka自带的SMO类库，速度上会有很大提高，调整kNN算法距离权值方法，原本是默认的0，现在调整为1-distance。
8. 协助刘卓然对分类器训练以及分类器分类正式任务进行错误测试及调整。

一点感想

项目还差以下几点没有完成，全自动服务，各种语料库的建立和完善，尤其是文本倾向性分析语料库。目前分类器分类效果较差，原因预测有以下几个：1.分词器停用词几乎没有，许多杂词对语料干扰太大。2.分类器细节参数设置不准确，API调用不熟，需要改进。

这次项目对以后工作有很多经验教训可以吸收，首先是采集器阶段，工作量大，内容多，难度一般，需要多人开发，项目后期发现，之前编写的采集器源程序出现较多bug，仅一个人了解核心程序难以发现潜在问题；第二，整个项目文件系统设计复杂，难以维护，项目持久层设计磁盘文件，mysql以及NoSql数据库MongoDB，关系错综复杂，需要简化；第三，核心算法模块过分依赖开源工具，开发人员对于核心算法及其使用意义理解不够深入，API使用不熟练；第四，选用核心开发工具封装化过于严重，开发自由度低，今后项目建议使用机器学习工具而非NLP工具，这样，虽然代码量增加，需要自己准备数据模型，但是可以根据系统功能做特定的数据准备，系统运行效果更加，也更有助于理论研究。

下一阶段准备和实验室同学一起继续努力，同时自己的注意力将更专注于以上问题的解决和完善，尤其是第三第四点，项目部分内核功能需要项目组其他同学帮助完成，本人也会尽最大努力帮助其迅速掌握。希望为实验室做的更多，得到更多的批评，得到更多的信任。